

# Blue Matter: Parallel Biomolecular Simulation for Blue Gene

Blake G. Fitch

Biomolecular Dynamics and Scalable Modeling

Thomas J. Watson Research Center

IBM, Yorktown Heights, NY 10598

<http://www.research.ibm.com/bluegene/>

October 14, 2003

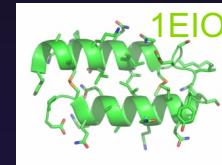
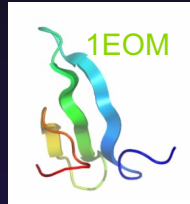
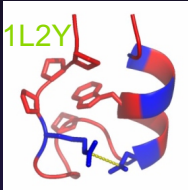
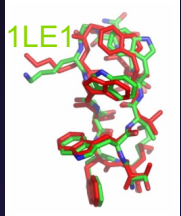
# Outline

- Introduction
- Parallelization of molecular dynamics
- Blue Matter overview
- Decomposition explorations and crude performance models
- 3D-FFT for BG/L (early measurements on hardware)
- Status

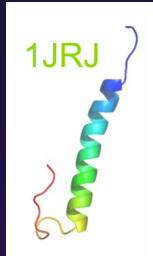
# Current Status

- Blue Matter molecular simulation application has run and regressed on early BG/L hardware (leverages extensive testing infrastructure)
- Developed implementation of “volumetric” decomposition of 3D-FFT as an alternative to typical “slab” decomposition to enable scaling to large numbers of nodes. Implementation can leverage any uniprocessor 1D-FFT implementation.
- Early results from MPI version of 3D-FFT on large SP (Power4) cluster show scaling superior to that of FFTW on the same platform
- MPI-based implementation of 3D-FFT and Blue Matter now running on production software stack (up to 512 node BG/L)
- Active message-based implementation of 3D-FFT and Blue Matter now running on multi-chip BG/L hardware using both cores
- Tuning and experimentation is just beginning

# Blue Gene: A spectrum of possible projects

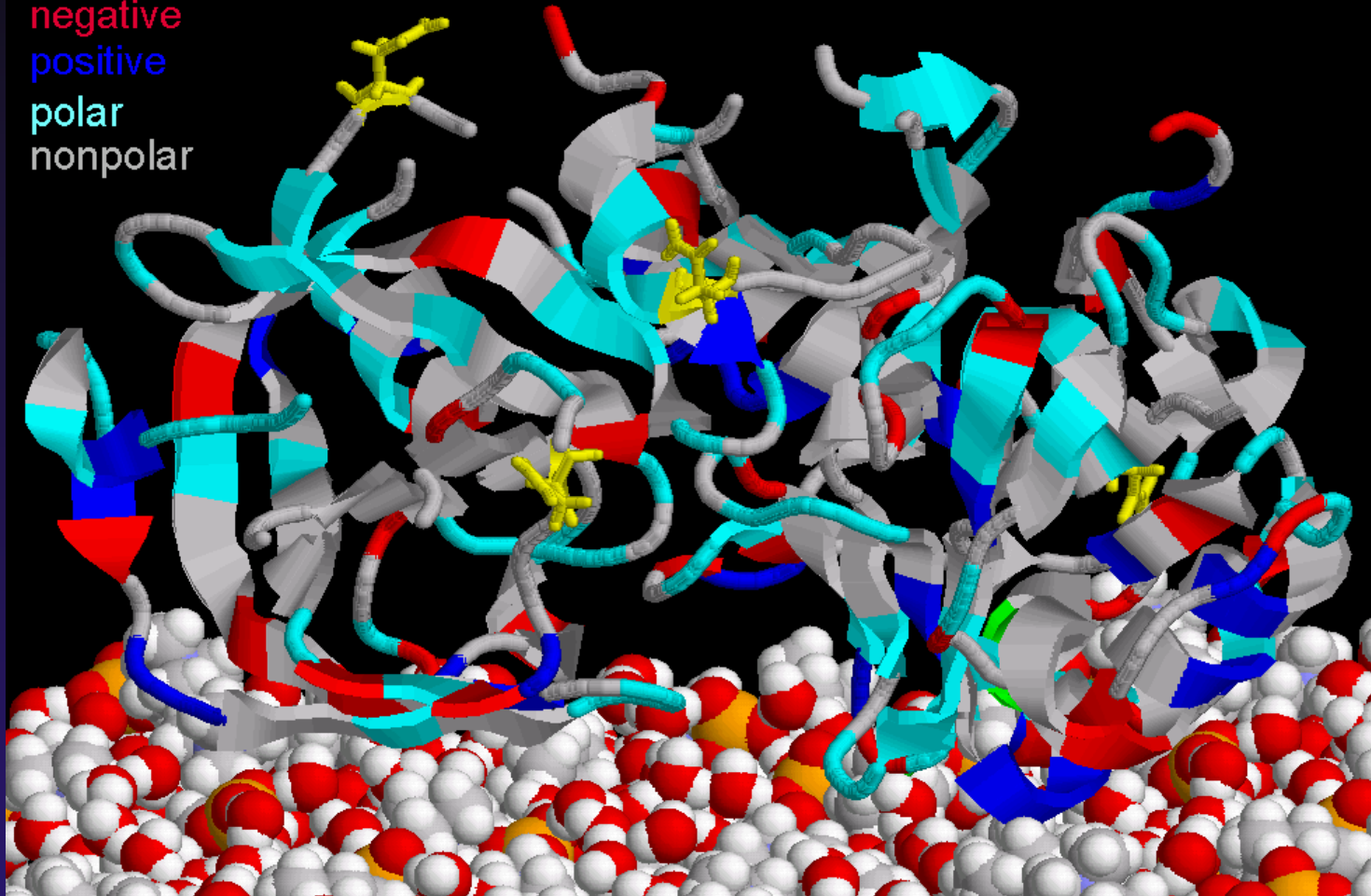


- cover a range of system sizes, topological complexity
- cover a broad range of scientific questions and impact areas:
  - thermodynamics
  - folding kinetics
  - membranes and membrane-bound systems
  - folding-related disease (CF, Alzheimer's, BSE)
- improve our understanding not just of protein folding but protein function



# beta-Secretase

negative  
positive  
polar  
nonpolar



# Design Points

- system size 10K-100K atoms with focus on 30K atoms
- biomolecules
- explicit treatment of solvent
- extensive infrastructure for validation and binary regression
- 1K-100K nodes
- modular architecture:
  - cellular MPP for computational core
  - large AIX/SMP for analysis
  - DB2 for experimental metadata
- separation of parallel programming complexity from molecular modeling complexity

# What Limits the Scalability of MD?

- Inherent limitations on concurrency:
  - Bonded force evaluation?
    - \* Represents only small fraction of computation, can be distributed moderately well.
  - Real space non-bond force evaluation?
    - \* Large fraction of computation, but good distribution can be achieved using volume or interaction decompositions.
  - Reciprocal space contribution to force evaluation?
    - \* Most prevalent methods (particle-mesh) rely on convolution evaluated using 3D-FFT—FFT involves global communication.
- Load balancing.
- System software overheads (particularly communications software).

# Real Space Domain Decompositions

- atom

- commensurate with propagation of dynamics
- limited ability to distribute work
- potential to generate strange (global) communications patterns

- volume

- commensurate with propagation of dynamics, 3D mesh machine, and limited range forces
- load balancing issues
- “strict” volume decomposition may conflict with efficient implementation of rigid subunits

- interaction

- good load-balancing characteristics
- implementation (Plimpton) available with efficient use of 2D mesh topology
- incommensurate with propagation of dynamics

# What Is Blue Matter?

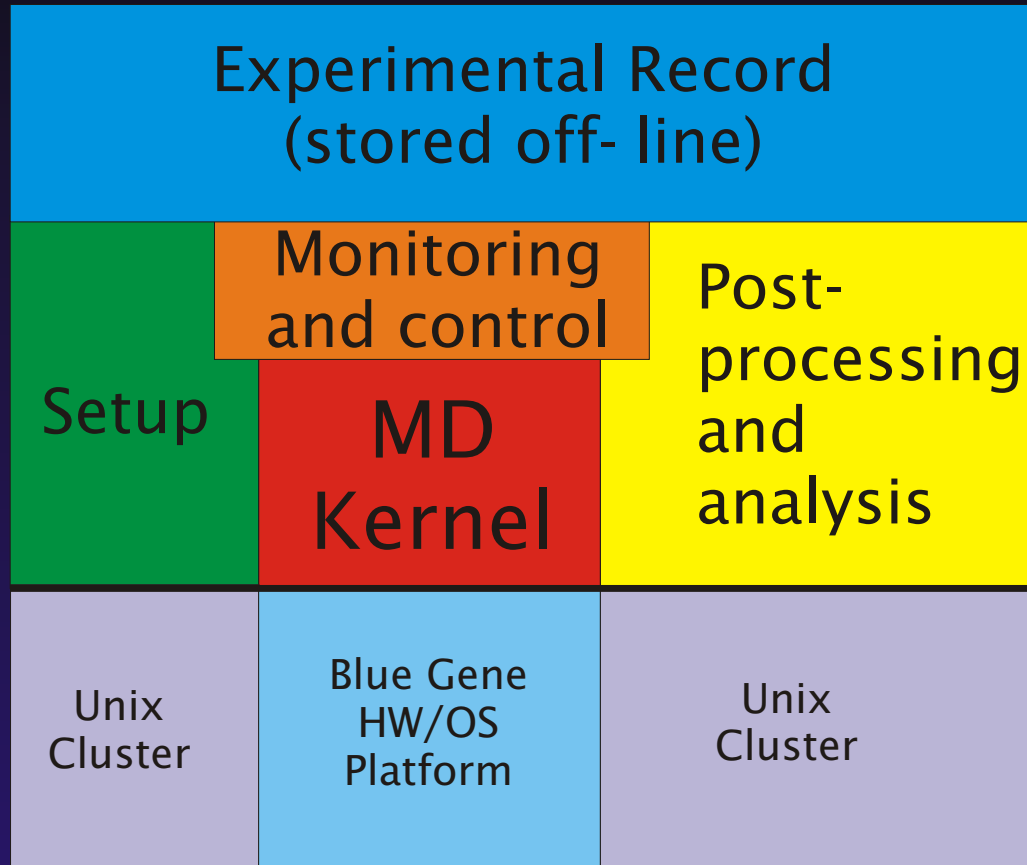
- Application platform for the Blue Gene Science program
- Prototyping platform for exploration of application frameworks suitable for cellular architecture machines
- Blue Matter comprises all of the necessary application components—those that run on the computational core and those that run on the host

For more details, see: **Fitch *et al.*, Blue Matter, An Application Framework for Molecular Simulation on Blue Gene**, Journal of Parallel and Distributed Computing Volume 63, Issues 7-8 July-August 2003 , Pages 759-773

# Blue Matter Overview

- Separate MD program into multiple subpackages (offload function to host where possible):
  - MD core engine (massively parallel, minimal in size)
  - Setup programs to setup force field assignments, etc
  - Monitoring and analysis tools to analyze MD trajectories, etc.
- Multiple Force Field Support
  - CHARMM force field (done)
  - OPLS-AA force field (done)
  - AMBER force field (done)
  - GROMOS force field (in progress)
  - Polarizable Force Field (planned)

# Application Overview

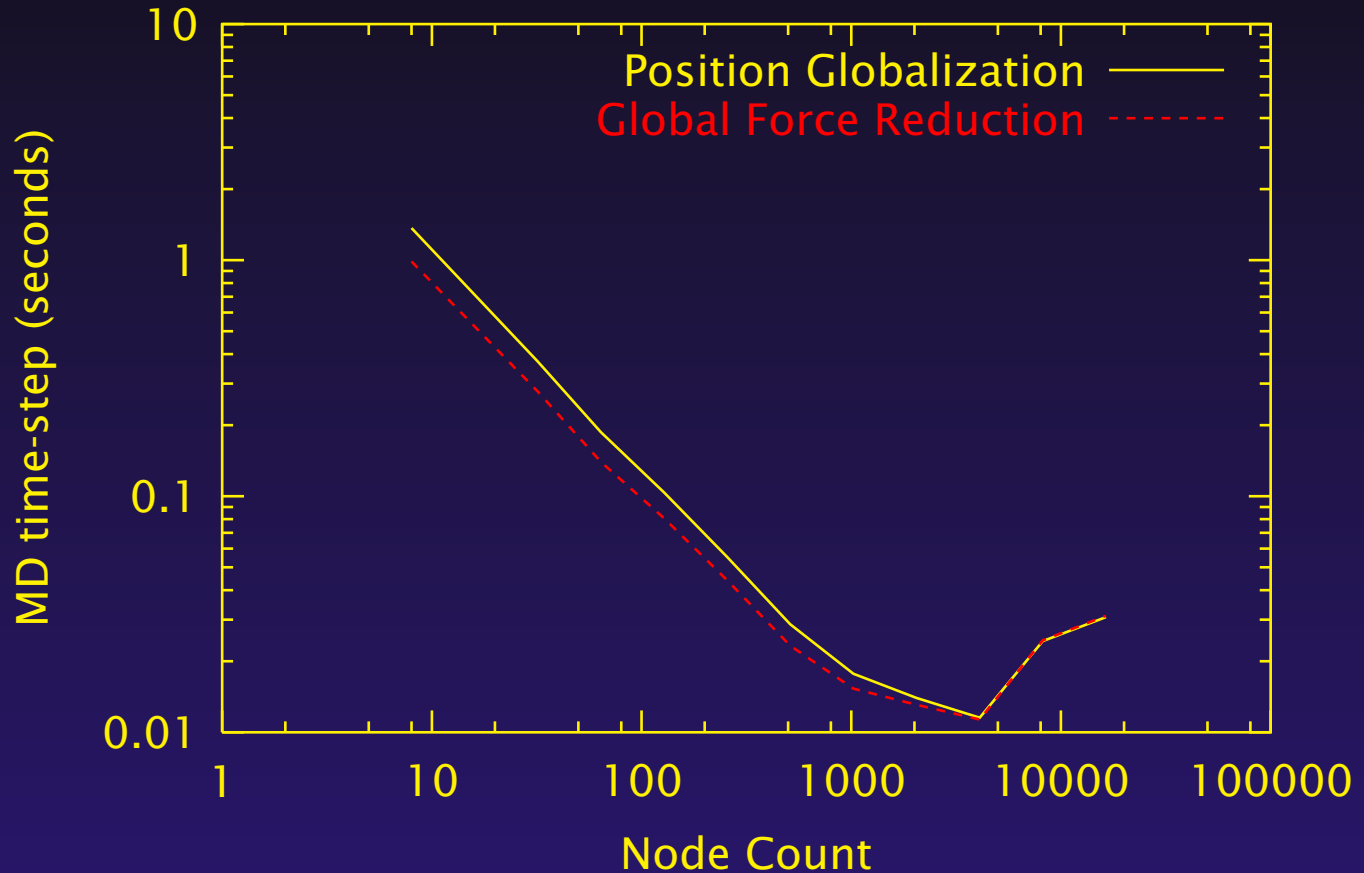




# Early Options for Parallel Decomposition Targeting BG/L

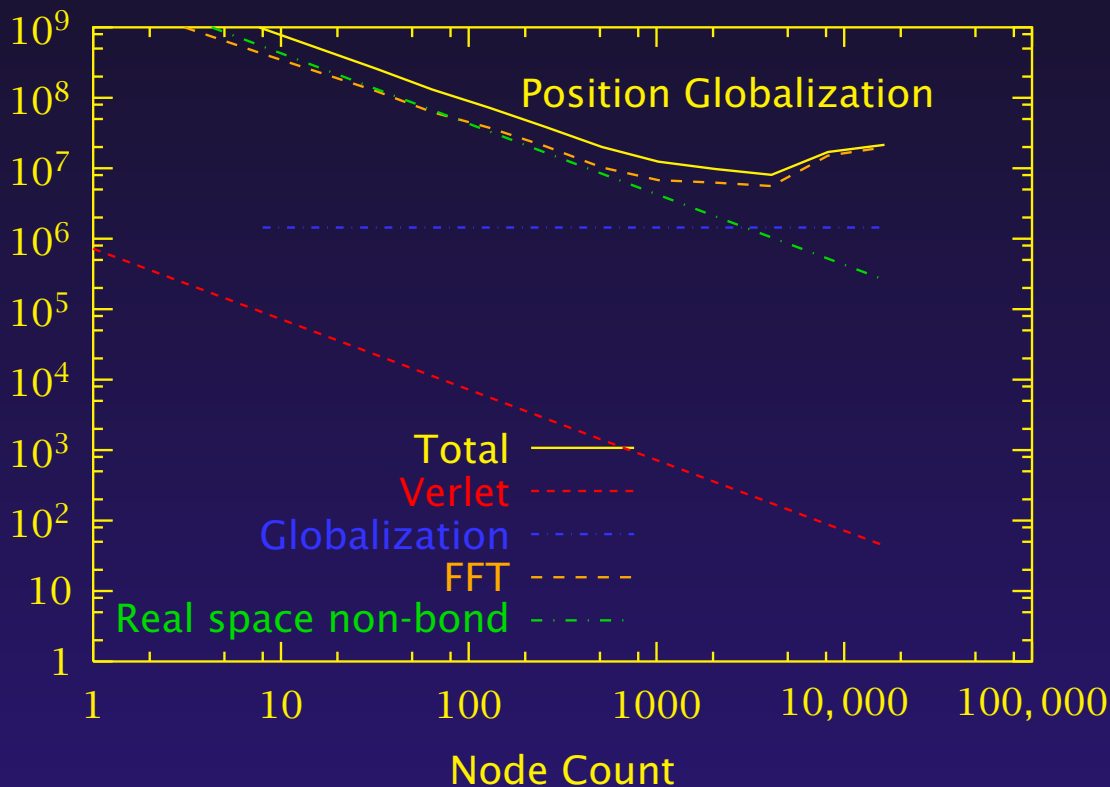
- Global force reduction — replication of dynamics propagation
- Globalize positions — double computation of forces, P3ME issues
- Globalize positions with nearest neighbor force reduction with approximate volume decomposition — supports P3ME
- Globalize positions with near neighbor (cutoff radius) force reduction with approximate volume decomposition — supports P3ME, avoids double computation of forces

# Molecular Dynamics, 30,000 Atoms



# Contributions to MD Time Step (Position Globalization)

Contributions to Time Step Duration (cycles)

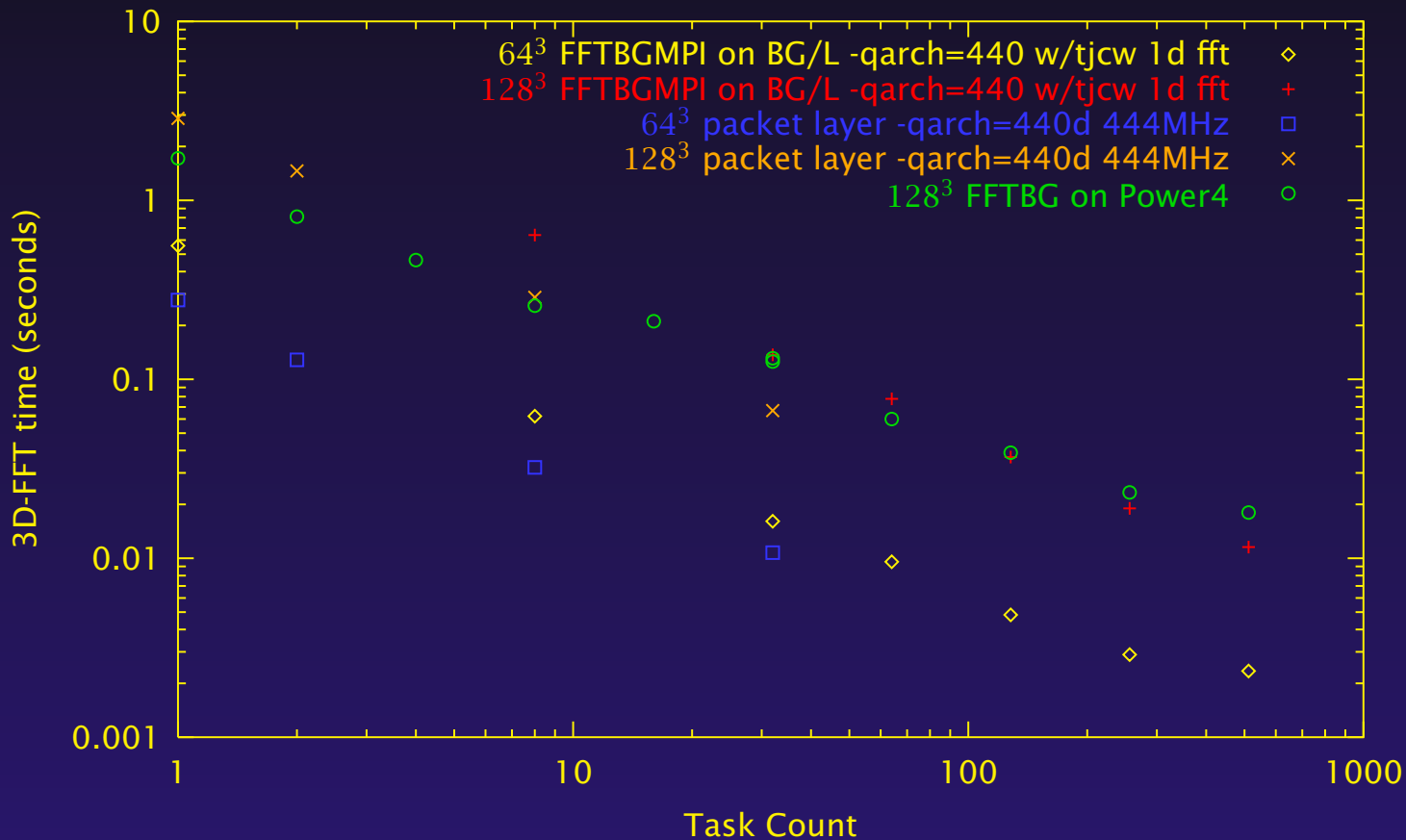


# 3D-FFT for BG/L

- **Requirement:** Scalable 3D FFT for meshes ranging from  $32^3$  to  $256^3$  as part of particle mesh molecular dynamics methods commonly used in biomolecular simulations.
- **Design goal:** "Volumetric" 3D FFT decomposition with strong scaling characteristics for mesh sizes of interest (as an alternative to "slab" based 3D FFT decomposition).
- **Sizing:** Paper and pencil sizing indicates that mesh sizes of interest will scale to thousands of nodes.
- **Prototyping:** "Active Packet" and MPI versions of volumetric 3D FFT have been implemented.
- **Results:** MPI version shows scaling on SP (Power4) superior to that of FFTW; MPI version on BG/L has run up to 512 nodes; Active Packet version running in BG/L bringup/test environment and scales well to 32 BG/L nodes.

Eleftheriou *et al.*, A Volumetric FFT for Blue Gene/L, to appear in the Proceedings of HiPC2003

# Volumetric 3D FFT on SP (Power 4) and BG/L



# 1D-FFT on BG/L (3D-FFT Building Block)

- Naive 1D-FFT implemented for 64 and 128 points (2-element vector implementation to take advantage of hummer<sup>2</sup>)
- Using IBM xLC compiler on Power3 (-O3 -qarch=pwr3) and BG/L hummer<sup>2</sup> (-O3 -qarch=440d)
  - 6643 cycles per 128 point FFT pair on Power3 (34% of peak)
  - 7346 cycles per 128 FFT pair on BG/L (30% of peak)
- Better performance should be expected from more sophisticated 1D-FFT implementations (such as the work being done at the Technical University of Vienna)

# BG/L Dual Core 3D FFT Flowchart

Core 0



Core 1



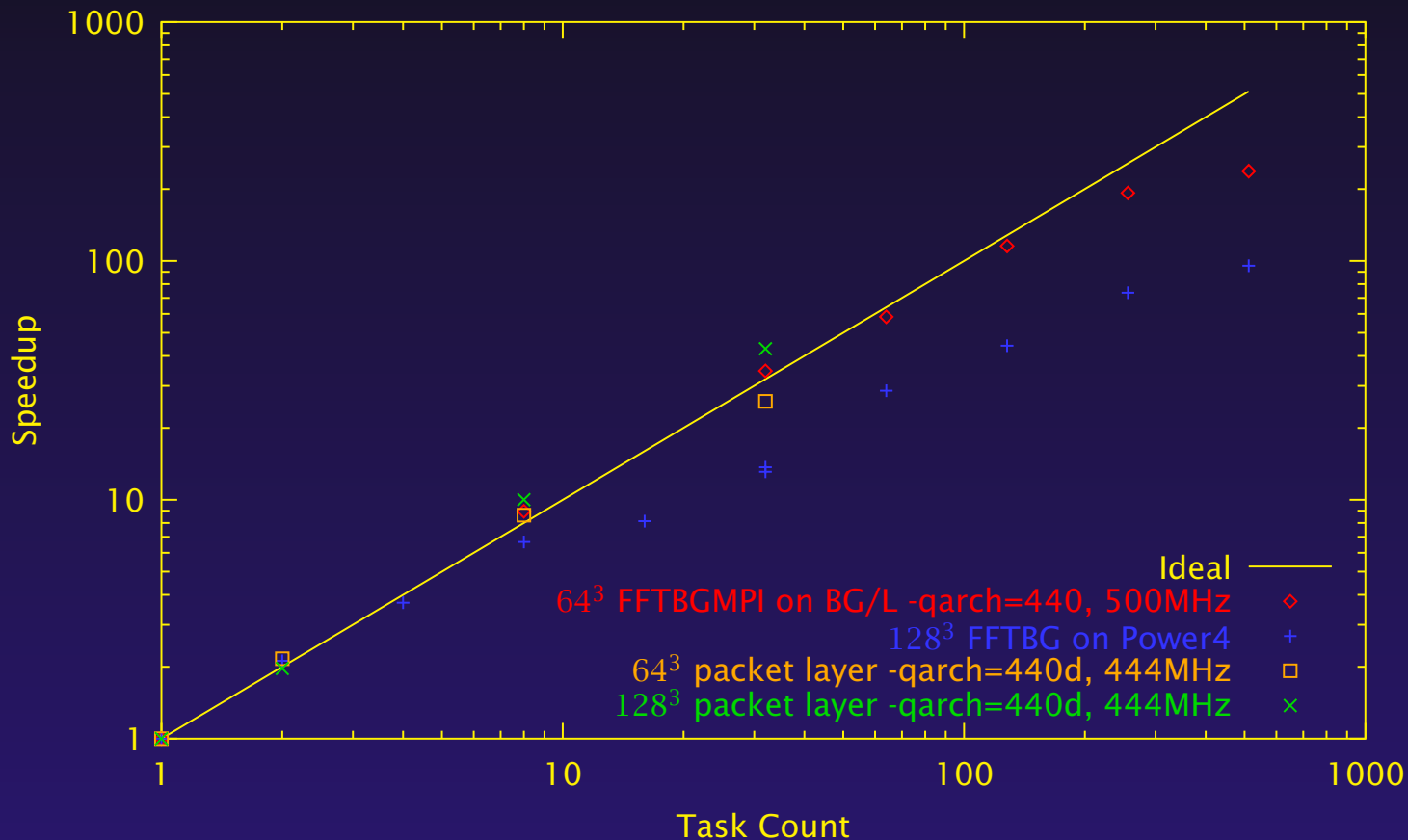
Time --- ►

Green : Active packet send in row or plane.

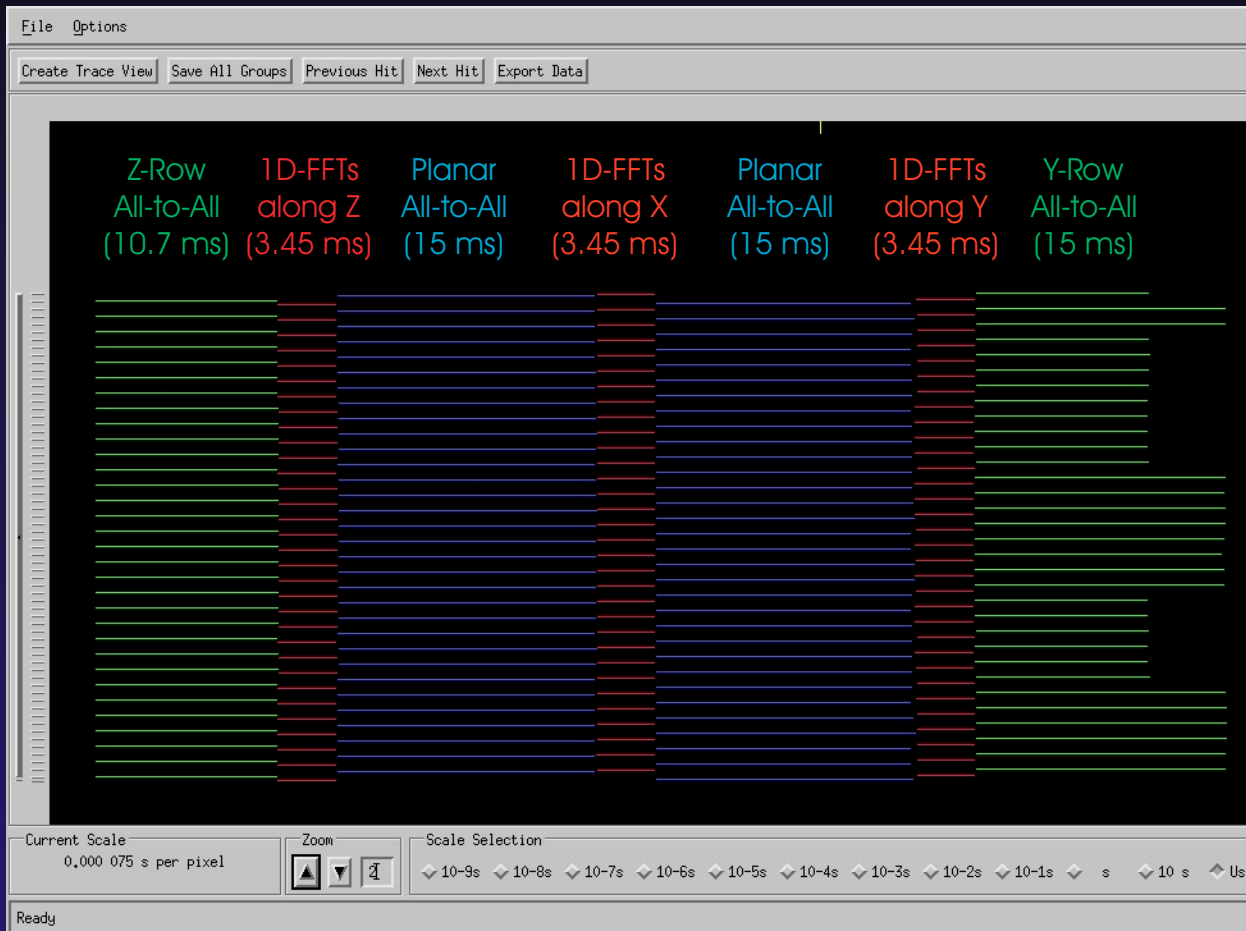
Blue : Active packet receive and dispatch (from BG/L FIFOs and internal Queue ).

Red : Compute 1/2 nodes allocation of 1D FFTs in phase

# Measured 3D FFT Speedups on BG/L and SP (Power4)



# 32 Node Application Tracing

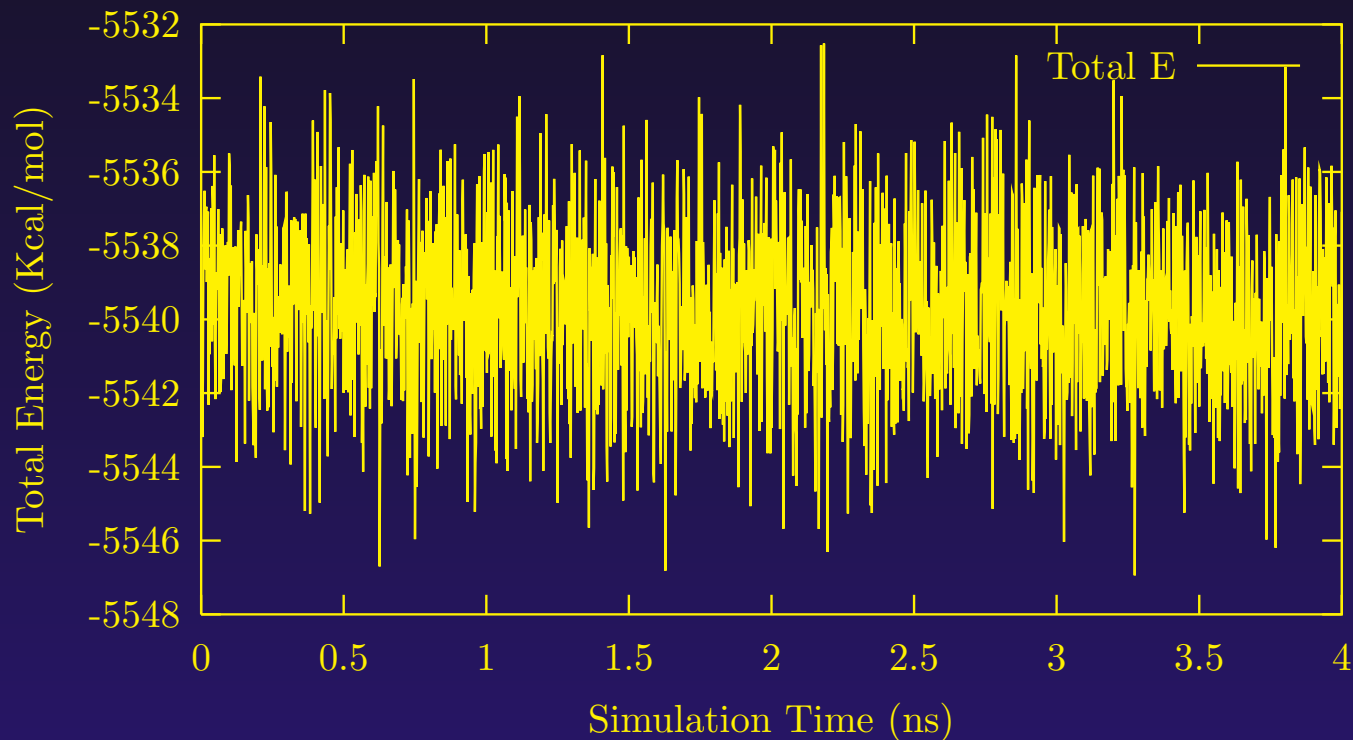


# Regression/Validation

- External regression of multiple force fields intramolecular energies via “Brooks suite”
- Main check on validity is quadratic dependence of conserved quantity rmsd vs. timestep for multiple short runs for every combination of features
  - Necessary but not sufficient many times
  - Agreement of p3me with Ewald in limit of small grid spacing
  - Good temperature and pressure behavior
- Testing “corner” cases ( e.g. Imaging )
- Binary regression capability is needed due to the experimental hw/sw environment

# 4ns NVE Lipid Run using Blue Matter

Total Energy for Fully Hydrated SOPE



# Conclusions from Simple Models

- Raw hardware “speeds and feeds” allow the use of standard MD techniques (P3ME) to node counts in excess of 1000 for problem sizes of interest to the Blue Gene project.
  - No extraordinary algorithmic methods have to be employed based on hardware considerations.
- “Naive” parallelization strategies leveraging global tree network are practical for a reasonable range of node counts even though the communication time is not scalable because:
- Scalability is limited by 3D-FFT—hardware latencies dominate for large node counts on fixed size problems as message sizes decrease and messages sent/node increase.

# Current Status

- Blue Matter molecular simulation application has run and regressed on early BG/L hardware (leverages extensive testing infrastructure)
- Developed implementation of “volumetric” decomposition of 3D-FFT as an alternative to typical “slab” decomposition to enable scaling to large numbers of nodes. Implementation can leverage any uniprocessor 1D-FFT implementation.
- Early results from MPI version of 3D-FFT on large SP (Power4) cluster show scaling superior to that of FFTW on the same platform
- MPI-based implementation of 3D-FFT and Blue Matter now running on production software stack (up to 512 node BG/L)
- Active message-based implementation of 3D-FFT and Blue Matter now running on multi-chip BG/L hardware using both cores
- Tuning and experimentation is just beginning

# Acknowledgements

## Blue Gene Application/Science team:

Alex Rayshubskiy, Blake Fitch, Maria Eleftheriou, Jed Pitera, Mike Pitman, Frank Suits, Bill Swope, Chris Ward, Yuri Zhestkov, Ruhong Zhou

## Blue Gene Hardware and System Software teams, particularly:

Mark Giampapa, Alan Gara, Philip Heidelberger, Burkhardt Steinmacher-Burow, Mark Mendell

## Collaborators:

Bruce Berne (Columbia), Scott Feller (Wabash), Klaus Gawrisch (NIH), Vijay Pande (Stanford)